Towards accurate differential diagnosis with large language models

Year: 2023 | *Citations: 194* | *Authors: Daniel McDuff, Mike Schaekermann, Tao Tu*

Abstract

A comprehensive differential diagnosis is a cornerstone of medical care that is often reached through an iterative process of interpretation that combines clinical history, physical examination, investigations and procedures. Interactive interfaces powered by large language models present new opportunities to assist and automate aspects of this process1. Here we introduce the Articulate Medical Intelligence Explorer (AMIE), a large language model that is optimized for diagnostic reasoning, and evaluate its ability to generate a differential diagnosis alone or as an aid to clinicians. Twenty clinicians evaluated 302 challenging, real-world medical cases sourced from published case reports. Each case report was read by two clinicians, who were randomized to one of two assistive conditions: assistance from search engines and standard medical resources; or assistance from AMIE in addition to these tools. All clinicians provided a baseline, unassisted differential diagnosis prior to using the respective assistive tools. AMIE exhibited standalone performance that exceeded that of unassisted clinicians (top-10 accuracy 59.1% versus 33.6%, P = 0.04). Comparing the two assisted study arms, the differential diagnosis quality score was higher for clinicians assisted by AMIE (top-10 accuracy 51.7%) compared with clinicians without its assistance (36.1%; McNemar's test: 45.7, P < 0.01) and clinicians with search (44.4%; McNemar's test: 4.75, P = 0.03). Further, clinicians assisted by AMIE arrived at more comprehensive differential lists than those without assistance from AMIE. Our study suggests that AMIE has potential to improve clinicians' diagnostic reasoning and accuracy in challenging cases, meriting further real-world evaluation for its ability to empower physicians and widen patients' access to specialist-level expertise. Diagnostic reasoning using an optimized large language model with a dataset comprising real-world medical cases exhibited improved differential diagnostic performance as an assistive tool for clinicians over search engines and standard medical resources.